Linear Regression (Revisited): Multiple Regression

STAT 245

Sept. 12, 2024

Multiple regression

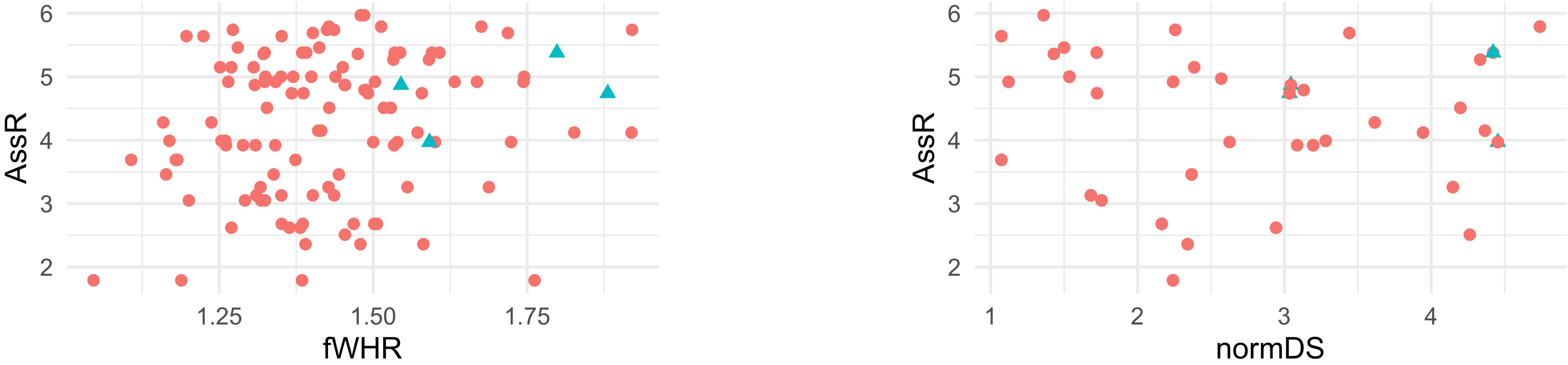
- Rarely does our response variable **really** depend on only one predictor.
- Can we expand our formulation to include more predictors? (Example: `normDS` also predicts `AssR`?)
- In R, it's super easy:

```
m2_2q <- lm(AssR ~ fWHR + normDS,
            data = bonobos)
```

Summary + Equation

```
## 
## Call:
## lm(formula = AssR ~ fWHR + normDS, data = bonobos)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.9993 -0.7592  0.1832  0.8279  1.7172 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  2.53889    0.85610   2.966  0.00369 **
## fWHR         1.40331    0.62298   2.253  0.02622 * 
## normDS      -0.09918    0.09687  -1.024  0.30810   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.094 on 113 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.04403,    Adjusted R-squared:  0.02711 
## F-statistic: 2.602 on 2 and 113 DF,  p-value: 0.07855
```

Prediction Practice



Prediction Practice

Show 10 entries

Search:

	fWHR	AssR	normDS
1	1.880866426	4.74	3.035
2	1.798387097	5.38	4.421
3	1.591439689	3.97	4.453
4	1.545018647	4.87	3.044

Showing 1 to 4 of 4 entries

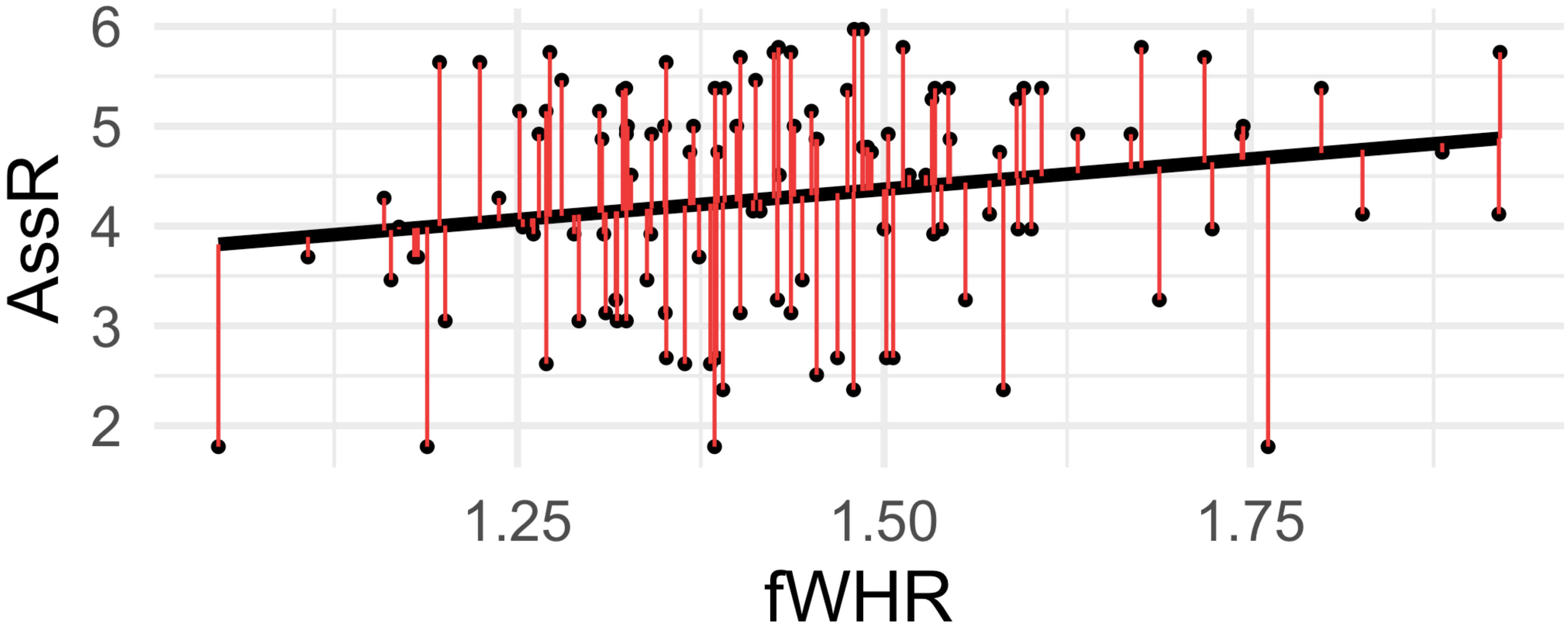
Previous 1 Next

Choosing Predictors, Again

- Here: build simple -> complex *to show math machinery*
- In practice: **Think before you model**
  - Rule of thumb (from Harrell): $p < \frac{n}{15}$
  - p is number of parameters want to estimate; n is sample size (rows in data)

How Fitting Happened

Simple Linear Regression Residuals



Least Squares Estimation

Minimize:

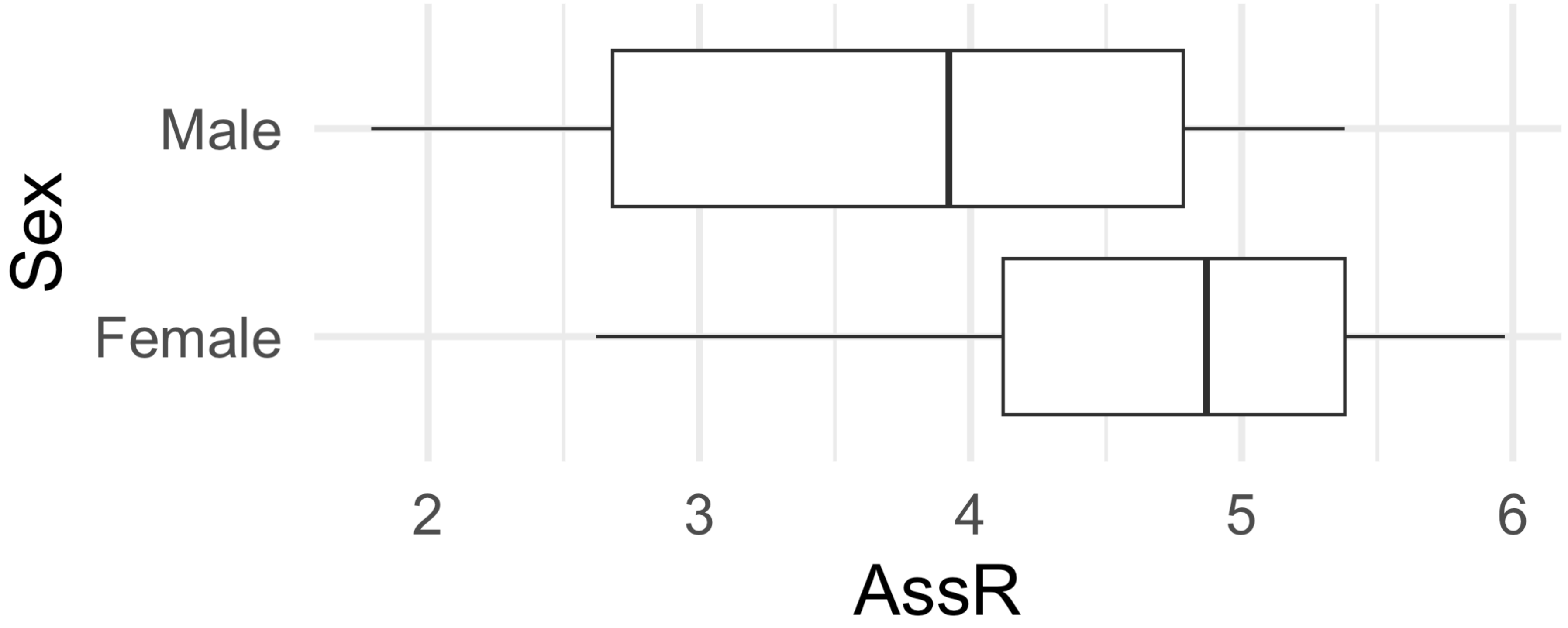
$$SSE = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Multiple Predictors?

- Harder to draw
- Just as easy to compute $\hat{y}$ …
- and thus compute the observed residuals e_i
- and the sum of squared residuals

See: https://setosa.io/ev/ordinary-least-squares-regression/

Predictors with 2 categories



Predictors with 2 categories

```r
m3_2q1b <- lm(AssR ~ fWHR + normDS + Sex,
           data = bonobos)
coef(m3_2q1b)
```

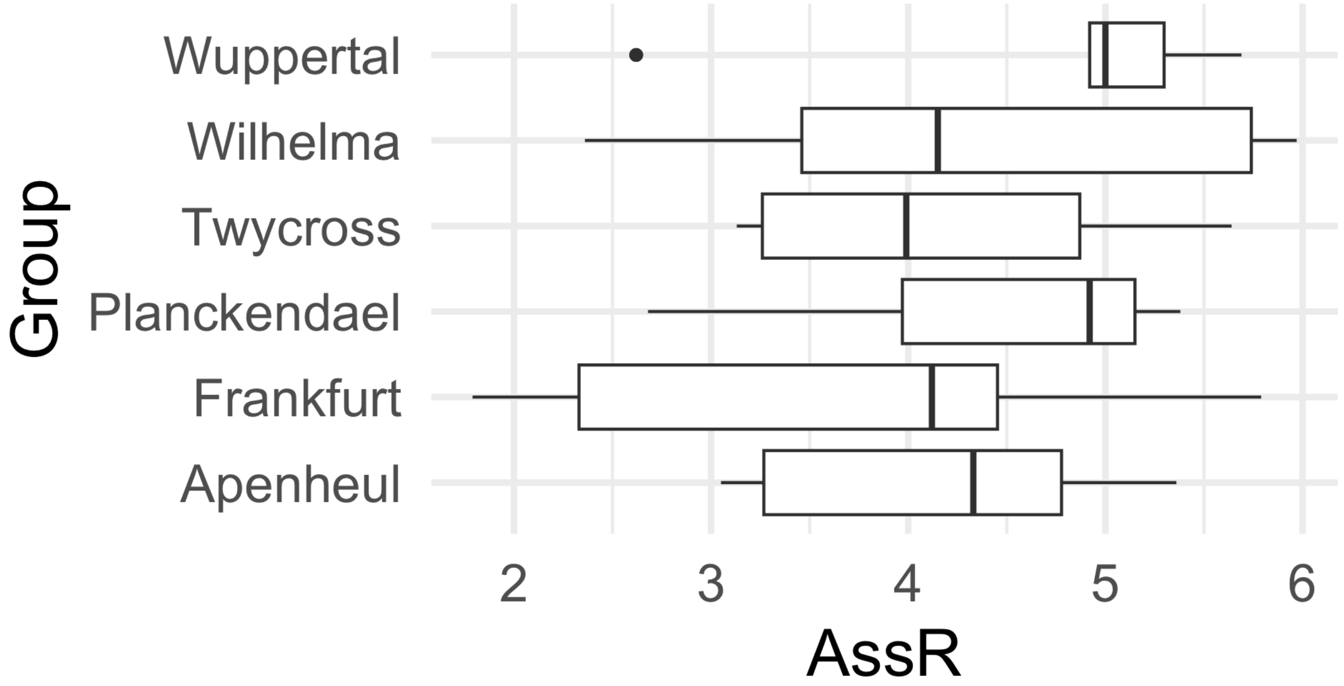
```
## (Intercept)        fWHR      normDS     SexMale
##  2.07913144  1.89581129 -0.01849396 -1.11030054
```

Predictors with 2 categories - Summary, Equation

```
##
## Call:
## lm(formula = AssR ~ fWHR + normDS + Sex, data = bonobos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.47788 -0.61852  0.09069  0.73386  1.59519
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.07913    0.75375   2.758 0.006786 **
## fWHR         1.89581    0.55186   3.435 0.000831 ***
## normDS      -0.01849    0.08592  -0.215 0.829962
## SexMale     -1.11030    0.18681  -5.943 3.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9581 on 112 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2732,    Adjusted R-squared:  0.2538
```

More categories

```
gf_boxplot(Group ~ AssR,
           data = bonobos)
```



More Categories

```
m3_2q2c <- lm(AssR ~ fWHR + normDS + Sex + Group,
           data = bonobos)
```

More Categories: Summary, Equation

```
##
## Call:
## lm(formula = AssR ~ fWHR + normDS + Sex + Group, data = bonobos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5148 -0.5901 -0.0118  0.6610  1.5405
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.16779    0.79783   2.717  0.00768 **
## fWHR                1.65461    0.56485   2.929  0.00415 **
## normDS              0.07067    0.08782   0.805  0.42277
## SexMale            -1.23398    0.18576  -6.643 1.32e-09 ***
## GroupFrankfurt     -0.61604    0.34951  -1.763  0.08083 .
## GroupPlanckendael   0.35141    0.31958   1.100  0.27398
## GroupTwycross       0.09313    0.30547   0.305  0.76105
## GroupWilhelma      -0.08112    0.33549  -0.242  0.80940
## GroupWuppertal      0.47304    0.32545   1.453  0.14901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Predictions by Hand

What is the expected `AssR` (according to this model) for 30 kg female bonobos at the Wilhelma zoo with `fWHR` of 1.5 and `normDS` of 2.5?

Predictions in R

Caution: missing data

```
bonobos <- bonobos |>
  mutate(preds = predict(m3_2q2c))
```

```
## Error in `mutate()`:
## ℹ In argument: `preds = predict(m3_2q2c)`.
## Caused by error:
## ! `preds` must be size 117 or 1, not 116.
```

Predictions in R

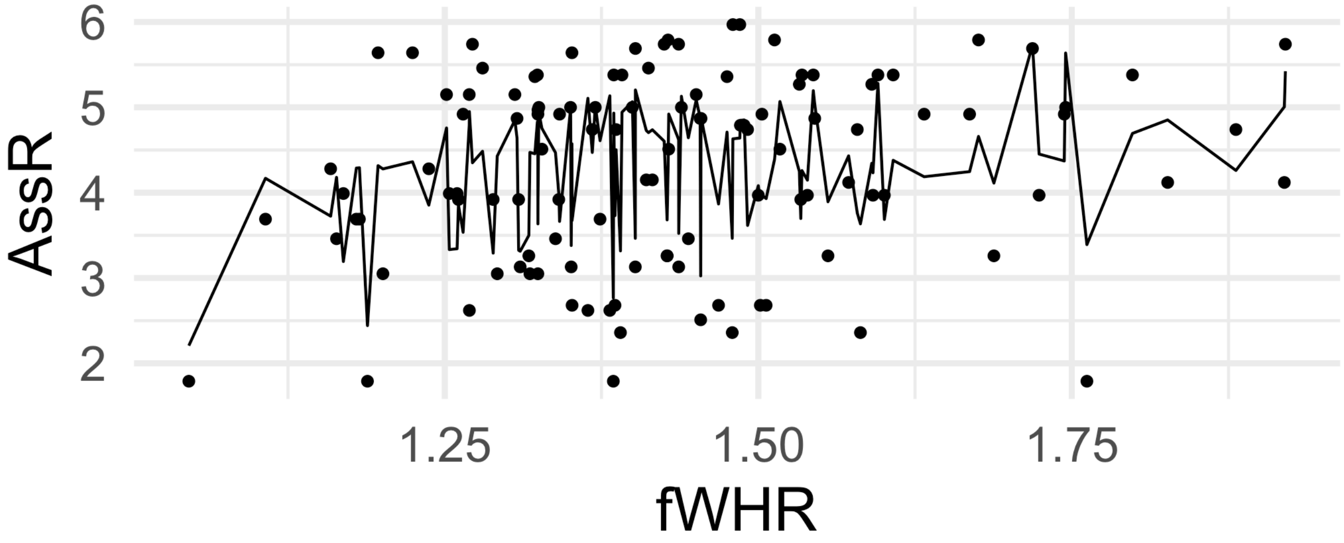
For ALL data points *in model*

```
b2 <- bonobos |>
  select(fWHR, normDS, AssR, Sex, Group) |>
  na.omit() |>
  mutate(preds = predict(m3_2q2c))
```

Plotting Predictions

Uh-oh, SO USELESS. Why should you *never do this?*

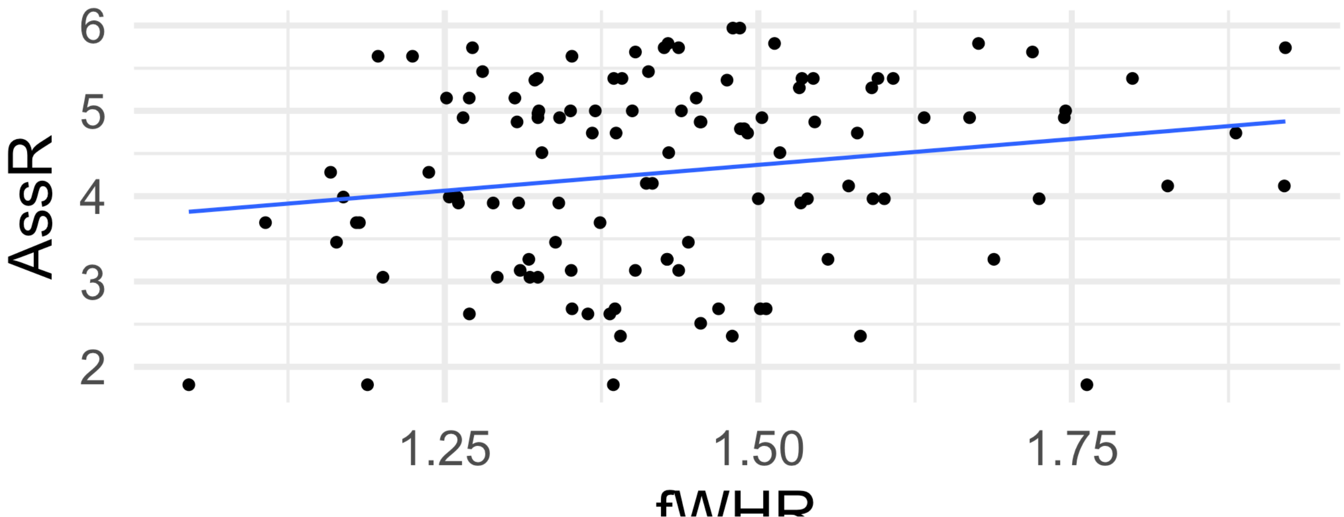
```
gf_point(AssR ~ fWHR,
          data = b2) |>
  gf_line(preds ~ fWHR)
```



`gf_lm()`: NEVER Do This Either

well, almost never -- why?

```
gf_point(AssR ~ fWHR, data = b2) |>
  gf_lm()
```


AssR
fWHR